Summarize to learn: summarization and visualization of text for ubiquitous learning

Rocio Chongtay
The University of Southern Denmark
Kolding, Denmark
rocio@sdu.dk

Mark Last
Ben Gurion University of the Negev
Beer-Sheva, Israel
mlast@bgu.ac.il

Mathias Verbeke
KU Leuven
Leuven, Belgium
mathias.verbeke@cs.kuleuven.be

Bettina Berendt
KU Leuven
Leuven, Belgium
bettina.berendt@cs.kuleuven.be

ABSTRACT
Visualizations can stand in many relations to texts – and, as research into learning with pictures has shown, they can become particularly valuable when they transform the contents of the text (rather than just duplicate its message or structure it). But what kinds of transformations can be particularly helpful in the learning process? In this paper, we argue that interacting with, and creating, summaries of texts is a key transformation technique, and we investigate how textual and graphical summarization approaches, as well as automatic and manual summarization, can complement one another to support effective learning.

Author Keywords
Text summarization, text visualization, ubiquitous learning, semantic enrichment.

ACM Classification Keywords
1.2.7 Natural Language Processing and text analysis; 1.2.6 Learning.

INTRODUCTION

Research reports indicate that text visualization, consistently improve learning performance [Levin, Anglin & Carney, 1987; Carney & Levin, 2002]. They suggest that text visualization becomes particularly valuable when the text transformation into a visual representation includes the recoding of information to make it more concrete and meaningful. In this paper, we argue that the process of creating and using text summaries is a key transformation technique for text visualization.

There are also research studies that have shown that people learn better with both words and pictures than with one or the other alone [Mayer, 2009]. In this position paper we present a design proposal for the MDIGESTS tool, which combines textual and graphical summarization approaches to support ubiquitous learning activities. Ubiquitous learning is frequently defined in terms of the underlying technologies [Cope and Kalantzis 2009]; our approach takes into consideration an instructional design with focus on the learner experience with summarization and visualization of text.

In this paper, we investigate the use case of technology-enhanced learning, in a combination of mobile and desktop settings.

Our key ideas are the following:

- We focus on learning approaches based on ubiquitous technology. This allows flexible access to learning resources at anytime from anywhere and with didactic constructs and tools that promote students' more active role in learning activities.
- We are fortunate that there is so much information available and constantly growing on the internet. However, this also leads to information overload. The summarization for learning approach presented here, attempts to streamline the focus on the specific learning topics to support effective ubiquitous learning.
- To reduce information overload, we employ a combination of Natural Language Processing and Interactive Systems, leveraging automatic summarization, visualizations, and rich interaction. We iteratively enhance this further by techniques such as semantic enrichment with the help of Linked Open Data.

In our approach, various summarization formats and learner activities complement one another. In each, the learner interacts with textual content, visualizations, and summarization tools in different ways.

USE CASE

Students from a course in "Knowledge Representation and Reasoning" are going to take a midterm exam on models of intelligence and the Web. These students have several exams to study for, and they could use their mobile devices to access information at anytime from anywhere to maximize their study time. However, reading full articles from a search on, for example, "machine intelligence" and selecting which are relevant is not very

efficient in this setting. The students use the MDIGESTS[1] tool which provides summarized descriptions of search results on the topic of interest from which the relevant articles can be rapidly selected and saved for later use as flashcards or as study material references.

MDIGESTS presents an interactive graphical automatic summarization of one long document or multiple documents, in which nodes represent relevant concepts and links represent some key relationship between the two nodes, typically a measure of co-occurrence frequency within the texts (see Figures 1 and 2).

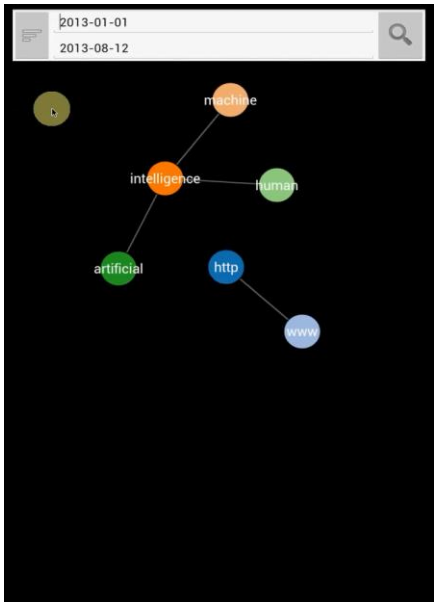

Figure 1. MDIGESTS screenshot: basic summary.

MDIGESTS rests on the single-document and multi-document summarization approaches of MUSE and its extensions [Litvak & Last, 2012] [Mirchev and Last, 2014] and of STORIES [Subašić & Berendt, 2010a, 2010b] and its mobile version MSTORIES [Aguilar, 2013].

The MUSE (MUltilingual Sentence Extractor) approach to single-document summarization [Litvak & Last, 2012] uses a linear combination of 31 language-independent features from various categories for ranking each sentence in a document. Language-independent features do not require any morphological or syntactic analysis of the summarized text, and they may include the sentence position in a document, the number of characters and words in a sentence, the similarity of the sentence to the document title, and other statistical metrics. MUSE finds the best set of feature weights by a genetic algorithm trained on a collection of human-generated document summaries. Formally, the MUSE model for sentence scoring can be expressed by the following formula:

$$Score = \sum w_i \times r_i ,$$

where r_i is the value of i^{th} sentence feature and w_i is its weight in the linear combination.

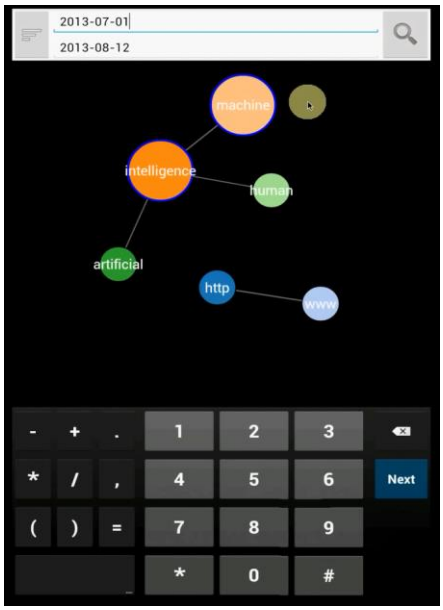

Figure 2. MDIGESTS Screenshot: Filtered summary (example filter document publication dates) and query specification (enlarged nodes).

The *summarization module* of MUSE performs an on-line summarization of input text(s). Each sentence of an input text document obtains a relevance score according to the trained model, and the top-ranked sentences are extracted to form the summary in their original order. To avoid duplicate content, a new sentence is added if and only if it is not similar to the previously selected sentences. The length of the resulting summaries is limited by a user-specified value (maximum number of words or sentences in the text extract or a maximum extract-to-text ratio).

The SentRel (Sentence Relations) multi-document summarization algorithm [Mirchev and Last, 2014] extends MUSE by assigning importance scores to documents and sentences in a collection based on two aspects: static and dynamic. In the static aspect, the significance scores are recursively inferred from a novel, tripartite graph representation of the text corpus. The graph includes three layers of disjoint entities: documents, sentences, and MUSE scores. In the dynamic aspect, the significance scores are continuously refined with respect to the current summary content.

STORIES was designed for summarizing the "stories" in time-indexed documents such as news articles, and its summarization approach is reused in MDIGESTS. For summarizing such documents, two strategies are important. First, articles can be filtered by the time frame

[1] MDIGESTS is a conceptual integration of the existing tools MSTORIES and MUSE (which are described below).

of interest. Second, the relevance of content is determined by how frequently it is mentioned across articles and relative to the entire time frame in which a story happens. Technically, STORIES rests on content-bearing terms (words and named entities weighted by measures such as TF.IDF) and their associations. For the summarization of a time-indexed subset of the whole corpus, c_t for period t, the frequency of the co-occurrence of all pairs of content-bearing terms b_j in documents is calculated as the number of occurrences of both terms in a window of w terms, divided by the number of all documents in c_t. We call this measure *local relevance* with $LR_t\ (b_1,b_2) = freq_t\ (b_1,b_2)$. LR normalised by its counterpart in the whole corpus C yields *time relevance* as the measure of burstiness: $TR_t\ (b_1,b_2) = (\ freq_t\ (b_1,b_2)\ /\ freq_C\ (b_1,b_2)\ )$. Thresholds are applied to avoid singular associations in small sub-corpora and to concentrate on those associations that are most characteristic of the period and most distinctive relative to others. This gives rise to the *story graphs* $G_t = <V_t\ ,E_t>$. The edges E_t are the story elements: all pairs (b_1,b_2) with absolute frequencies and TR above the respective thresholds. The nodes are the terms involved in at least one association in this symmetric graph: $V_t = \{\ b_j\ |\exists\ b_k : (b_j\ ,\ b_k\ )\ \exists\ E_t\ \}$.

In MDIGESTS, time generally plays a smaller role than in news summarization (although filtering by publication date is still relevant) and often, one long document rather than many small documents must be summarized. In this case, frequencies are aggregated over the whole document rather than over multiple documents. When there are documents or document parts of different relevance, the temporal processing of (M)STORIES is transformed into a processing based on the relevance weights, and burstiness is transformed into associations that are particularly salient in the relevant documents / document parts.

Based on the context of the text to be summarized, *semantic enrichment* can be used to add additional entities and links to the summary graph. This is done based on DBPedia, due to which this can be seen as extending the summary graph with the relevant part(s) of the semantic web graph. The DBpedia project [Bizer et al., 2009] extracts structured information from Wikipedia, and the DBPedia data contain entities that correspond to Wikipedia pages and relations between DBpedia entities and other resources, including other knowledge bases. In this way, not only the graph is enriched, but entities can also link out to relevant pages on Wikipedia. This offers the students additional (multilingual) study resources. This approach is similar to other semantic enrichment methods [Mendes et al., 2011; Milne and Witten, 2008], in particular Enrycher[2] [Štainer et al., 2010]. Our deep enrichment strategy (as described in [Lukovnikov, 2013]) is however optimized for short text fragments, and thus can also be leveraged for sentence level enrichment.

[2] http://ailab.ijs.si/tools/enrycher/

There are similar reports on web content summarization for mobile learning consumption, for example Yang et al (2012), however in that work, only textual summaries are retrieved and the users need to read them to select the relevant search results.

The main importance of the MDIGESTS visualization for learning is that it provides visualization with detail-and-context interaction. By selecting subgraphs, users can focus on parts of the graph that they are particularly interested in: students can use these graph parts for searching the most relevant content in a more efficient manner.

In the learning setting, this detail-and-context functionality is employed to give cues for testing one's knowledge: The learner (or the system) can select one edge (or a more complex subgraph) and test her knowledge of how these concepts relate to one another. An example in the scenario could be the link between ``machine" and ``intelligence".

First, the learner receives a multiple-choice test describing the semantics of the edge. If she answers this correctly, she can proceed to the next part of the graph. If she does not answer it correctly, or if she wants to know more about the concepts, she can use the edge (or subgraph) to query the document collection underlying the summarization. This digging deeper has two modes: In the summary mode (suited for the mobile learning setting), the learner receives an automatic, query-dependent summary of underlying documents, with the query being the chosen edge or subgraph. Thus, the graphical, STORIES-like, overview summary triggers a user interaction and user query, which lets the system respond with a more detailed sentence-based, MUSE-like, specific summary. This also answers often-voiced requests of STORIES users for sentential summary elements. (In the original STORIES news summarization, this request was met by a timeline depiction of highest-ranking sentences per date.) In the full-text mode (suited for the desktop learning setting), the learner can inspect the various documents that gave rise to the edge or subgraph.

These activities are largely receptive and should therefore be complemented by more constructionist elements in order to support deeper learning. This is done in the fourth component of our tool, the teaching of summary writing, which is designed for (although not restricted to) a desktop setting. The student reads the entire document or documents. The student starts writing her own abstractive summary with/without seeing the original text. The software limits the summary size (e.g., number of words). The software continuously calculates the scores of each summary sentence and of the entire summary. These scores can be presented continuously or on demand. The best and the worst summary sentences are presented in different colors. The student can add, modify

or remove summary sentences at any time subject to the summary size constraints. If relevant, the software presents sentences from the original texts that contain important information, which is missing in the student summary. The student can choose original sentences suggested by the software and then modify them. The calculation of scores and the system presentation of suggestions are based on the automatic summarization approaches described above.

CONCLUSION

We have discussed various summarization formats and their combination within a ubiquitous learning setting. Ubiquitous computing supports learning at anytime from anywhere. The use case presented in our position paper underlines that in order to achieve effective and in-depth learning, specific summarization formats and didactic design combinations should be applied. For example, in the mobile and desktop settings, the learner interacts with textual content, visualizations, and summarization tools in different and complementing ways. We have argued that not only the use of automated summaries can support the learning process, but that the writing of summaries can be trained by the use of automated text summarization tools. Transforming the text content into a summary requires the learner to recode its content to make it more concrete and meaningful, thus making summarization a key learning activity. As this position paper presents a design proposal of the MDIGESTS tool for summarization-based learning, future work will be focused on implementing, testing and exploring the potential of the proposed tool in a broader context on additional use cases.

REFERENCES

1. Aguilar, J., Mobile STORIES: Mobile news aggregator and interactive filtering, Master's thesis, KU Leuven, 2013.
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. DBpedia - A Crystallization Point for the Web of Data. Web Semant., vol. 7, no. 3, pp. 154–165, Sept. 2009.
3. Burbules, N., C. Meanings of Ubiquitous Learning. IN: Cope, B., and Kalantzis M., Eds. Ubiquitous Learning. University of Illinois Press. Urbana and Chicago, 2009.
4. Carney, R. N. and Levin, J. R. Pictorial Illustrations *Still* Improve Students' Learning From Text. Educational Psychological Review, vol. 14, no. 1, pp. 5-26, 2002.
5. Litvak, M., and Last, M. Cross-lingual training of summarization systems using annotated corpora in a foreign language. *Information Retrieval.* DOI http://dx.doi.org/10.1007/s10791-012-9210-3 , 2012.
6. Mayer, R.E. Multimedia Learning. Cambridge University Press, 2nd edition, 2009.
7. Mirchev, U. and Last, M. Multi-document Summarization by Extended Graph Text Representation and Importance Refinement. To appear in A. Fiori (Ed.), Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding, IGI Global, Hershey PA, USA. 2014.
8. Mendes, P., N., Jakob, M., García-Silva, A. and Bizer, C.. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pp. 1–8, New York, NY, USA, 2011. ACM.
9. Milne, D. and Witten, I. H. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM'08, pp. 509–518, New York, NY, USA, 2008. ACM.
10. Levin, J. R., Anglin, G. J., and Carney, R. N. On empirically validating functions of pictures in prose. In Willows, D. M., and Houghton, H. A. (eds.), The Psychology of Illustration: I. Basic Research, Springer, New York, pp. 51–85, 1987.
11. Lukovnikov, D., Tweenterest: Interest-driven topic-based tweet classification using semantic enrichment with DBpedia, Master's thesis, KU Leuven, 2013.
12. Štainer, T., Rusu, D., Dali, L., Fortuna, B., Mladeni´c, D., Grobelnik, M.: A service oriented framework for natural language text enrichment. Informatica. vol. 34, no. 3, pp 307-313, Ljubljana, 2010.
13. Subašić, I., & Berendt, B. Discovery of interactive graphs for understanding and searching time-indexed corpora. Knowledge and Information Systems, vol 23, no. 3, pp. 293-319, 2010a.
14. Subašić, I., Berendt, B. Experience STORIES: a visual news search and summarization system. Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, *Part III*. ECML/PKDD. Barcelona, September 20-24, 2010 (pp. 619-623) Springer, 2010b.
15. Yang G., Kinshuk, Sutinen E., and Wen D.: Chunking and Extracting Text Content for Mobile Learning: A Query-focused Summarizer Based on Relevance Language Model. In *Proceedings: IEEE International Conference on Advanced Learning Technologies*. pp. 126-128. Rome, Italy, 2012.